DOCUMENT RESUME

ED 371 036 TM 021 740

AUTHOR Bernstein, Lawrence; Burstein, Nancy
TITLE Bias vs. Precision: Combining Estimates in Multisite Evaluation Research.
PUB DATE Apr 94
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Bias; Comprehensive Programs; Data Analysis; Data Collection; Demonstration Programs; Error of Measurement; *Estimation (Mathematics); Evaluation Methods; Models; *Research Methodology; Research Problems; Statistical Analysis; Statistical Studies
IDENTIFIERS Comprehensive Child Development Program (ACYF); Hierarchical Linear Modeling; *Multiple Site Studies; *Precision (Mathematics); Weighting (Statistical)

ABSTRACT

The inherent methodological problem in conducting research at multiple sites is how to best derive an overall estimate of program impact across multiple sites, best being the estimate that minimizes the mean square error, that is, the square of the difference between the observed and true values. An empirical example illustrates the use of the following five models with data from the Comprehensive Child Development Program (CCDP), a 5-year national demonstration program implemented in 21 sites: (1) pooled data; (2) unweighted averaged; (3) weighted averaged; (4) hierarchical linear model random; and (5) hierarchical linear model fixed. Most striking is the similarity of results from all models. In the particular example, choice of model would not alter the conclusion that participation in the CCDP raised children's scores a given amount, although other outcomes might be more sensitive. By informing the analysis strategy with the employed sampling design, one can better justify the conclusions drawn regarding the efficacy of a particular program intervention. Two tables present analysis results. (Contains 13 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 371 036

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
LAWRENCE BERNSTEIN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Bias vs. Precision:

Combining Estimates in Multisite Evaluation Research

Lawrence Bernstein

Nancy Burstein

Abt Associates, Inc.
Cambridge, MA

Paper presented at the annual meeting of
the American Educational Research Association
New Orleans, LA

April 5, 1994

TM 021740

BEST COPY AVAILABLE

Introduction

Educational and social interventions are often implemented in multiple sites. By increasing the number of participants in a demonstration, a multisite study can increase the power for detecting significant differences between program groups. Moreover, such studies offer the prospect of replicating results with a variety of populations and in a variety of environments, which can enhance the external validity of a study in terms of the generalizability of its findings. Multisite demonstrations may be characterized by use of either identical protocols in all sites (typical of biomedical research) or of general program guidelines for the sites to implement as they see fit (typical of social science research). In both instances, multisite evaluations may be distinguished through their use of raw data from meta-analyses of programs, which synthesize summary statistical data such as standardized effect sizes from published research and evaluation studies (see Hedges, 1984). The inherent methodological problem in conducting research of this type lies in how best to derive an overall estimate of program impact across multiple sites.

By "best" we mean that estimate which minimizes the mean square error (MSE); that is, the square of the difference between the observed and true values. It can readily be shown that the MSE is the sum of the variance of the estimator and the square of the bias of the estimator.[1]

---

[1]Let y be the true value of the parameter that it is desired to estimate, $\hat{y}$ be the estimated value, and e the error of estimate. Then

$$e = \hat{y} - y = [\hat{y} - E(\hat{y})] + [E(\hat{y}) - y].$$

If we square this relationship and take expected values, we get:

$$MSE \equiv E(e^2) = E[\hat{y} - E(\hat{y})]^2 + [E(\hat{y}) - y]^2.$$

(continued...)

While statistical theory can provide guidance about variance, the bias of an estimator of the overall effect in a multisite evaluation also needs to be understood in the context of the real-world process by which sites were selected and the intervention was implemented as well. Hence choice of an optimal estimator of the overall effect will in the end depend on the judgment of the evaluator about these features of the experiment.

Review of the Literature

The methodological issues related to analyzing multisite evaluations have received scant attention in the evaluation literature. A recent volume of New Directions for Program Evaluation (Turpin & Sinacore, 1991) was devoted entirely to the issue of multisite evaluations. Despite the obvious benefits of studying implementations in multiple sites, the editors lament that "there is no available source for one to learn about this approach to evaluation" (1991, p.5).

The questions raised in this paper have been studied primarily within the biomedical community in the context of multiclinic trials. Fleiss (1986) noted two issues that need to be addressed when confronted with data from multiple sites. The first concerns how data should be pooled when deriving an estimate of an intervention. That is, should differences between treatment groups be averaged across sites, or should all the data be combined into a single analysis, ignoring site as a classification factor? The second issue, assuming that site effects are averaged, involves the optimal method for computing an overall impact estimate. Should each

---

[1](...continued)

It can easily be seen that the first term is the variance of $\hat{\gamma}$, while the second term is the square of the bias of $\hat{\gamma}$.

site be weighted equally, or should sites be weighted as a function of sample size or in some other way?

Although multisite evaluations have been conducted on numerous occasions in the past, little attention has been paid to the issues raised by Fleiss. One common approach has been to ignore differences among sites. An example of this approach is the National JTPA Study (Bloom et al., 1992), an evaluation of the Jobs Training Partnership Act of 1982, based on participant experiences in 16 different local service areas. The analysis bypassed the estimation of site-specific impacts, and instead pooled the research sample across all the sites to estimate a global mean impact of the program. Although perhaps intuitively appealing, this analytic approach has two drawbacks (Fleiss, 1986). First, because random assignment of individuals or families to program or comparison groups normally takes place at the individual site level, this approach is theoretically inappropriate. Second, this approach is statistically inefficient because the variance of the estimated combined impact may be inflated due to the differences among site means. This latter concern has arisen in the debate concerning the appropriate unit of analysis in large-scale evaluations such as Head Start and Follow Through (Haney, 1980).

Assuming that site differences should be accounted for in the analysis, a second issue is to determine the optimal method for averaging individual site estimates to yield an overall impact estimate. Once again, there is little guidance from the literature as to how best to proceed. Fleiss (1986) recommends weighting estimates as a function of sample size.[2] If a significant

---

[2]The site weights that are typically used to combine site-level impacts are:

$$w_s = \frac{n_{sp} \times n_{sc}}{n_{sp} + n_{sc}}$$

(continued...)

site by treatment effect is found to exist, however, then Fleiss recommends weighting each site's effect equally.

This criterion was adopted in the Infant Health and Development Program evaluation, a large multisite clinical trial testing the effectiveness of a combination of early childhood and family support services on reducing developmental and health problems in low birthweight infants (Infant Health and Development Program, 1990). As reported in this study, the hypothesis of homogeneous effects across site and birthweight groups was tested for each outcome of interest. If this hypothesis was not rejected, that is, if the effect of treatment was apparently the same across all site by weight subgroups, then the common effect size was estimated by averaging the subgroup effects weighted by a factor of $n_g * p_g * (1-p_g)$ where n_g was the sample size in subgroup g, and p_g represented the proportion in that subgroup assigned to the treatment group. If, on the other hand, the effects of the treatment were found to be different over the subgroups, an equal-weighting approach was adopted.

These issues have also been discussed within the context of the conduct of agricultural experiments. Cochran and Cox's classic work on Experimental Design (1957) devoted a chapter to a discussion of the analysis of the results of a series of experiments. In the classical analysis of variance framework, Cochran and Cox address the dual questions of determining whether the treatment effects are the same in all experiments, and of estimating the average treatment effect across a set of experimental replications. Although the methodological framework is different,

---

[2](...continued)
where n_{sp} and n_{sc} are the numbers of program participants and control group members, respectively, in site *s*. These weights take into account both differences in sample sizes across sites and within sites between program and comparison groups.

the same guidance is offered with regard to weighting. In the absence of interactions between treatment and place, weighting each treatment mean inversely proportional to its variance maximizes statistical power, with negligible cost in terms of bias. If a weighted analysis is conducted in the presence of interaction, however, the F-ratio for the main treatment effect is unduly biased upwards, producing too many significant results. In this case, an unweighted analysis is preferable.

A final contribution to the theory of multisite evaluation appears in the literature on multilevel or hierarchical linear modeling approaches (Bryk & Raudenbush, 1992; Raudenbush, 1988). Although the utility of these approaches is primarily for explaining within-group variance as a function of group characteristics, the statistical theory underlying these approaches also has applications to the issues under consideration here. In its simplest form, a model conceptually equivalent to a one-way random effects analysis of variance with groups (e.g., site) treated as a random factor can be written as follows:

$$Y_{ij} = \beta_{0j} + \epsilon_{ij} \qquad \text{(within-group) and} \qquad (1)$$

$$\beta_{0j} = \gamma_{00} + \upsilon_{0j} \qquad \text{(between-group)} \qquad (2)$$

In this formulation, model (1) states that the outcome variable Y_{ij} for individual i within group j varies around a group mean β_{0j} with independent errors ϵ_{ij} assumed to be distributed $\sim N(0,\sigma^2)$, where σ^2 signifies within-group or sampling variance. In model (2), each group's mean β_{0j} varies around a grand mean γ_{00} with independent errors υ_{0j} distributed $\sim N(0,\tau_{00})$ with τ_{00}

signifying between-group or parameter variance[3]. This model is particularly relevant in terms of the estimation of the grand mean γ_{00}. In the case of unequal sample sizes across groups, then a weighted least squares estimator of γ_{00} can be expressed as a precision-weighted average of the sample means $\overline{Y}_{.j}$ (estimators of the β_{0j}), where each sample mean is weighted inversely proportional to its variance. This estimator $\hat{\gamma}_{00}$, a unique minimum-variance unbiased estimator of γ_{00}, is reduced to a simple arithmetic average of the group sample means when all the precisions (reciprocal variances) are equal (Bryk & Raudenbush, 1992). If a within-group variable such as treatment group membership is included in the model, then the average slope measuring the impact of the treatment across groups, γ_{10}, can similarly be estimated. For a particularly relevant application of this procedure in assessing the impact of a decentralized technical and vocational educational initiative on Scottish secondary students, see Raffe, 1991.

Conceptualizing the Overall Effect

We believe that much of the confusion surrounding how to estimate the overall effect of an intervention stems from a lack of clarity about what is meant by the overall effect. Several alternative interpretations are possible, including:

- the average effect on those particular individuals who participated in a demonstration;

---

[3]The point estimates for the β_j are obtained under HLM using empirical Bayes procedures. The estimates are expressed as a weighted combination of the observed OLS estimates of the within-site means, and the estimated average weighted mean for the population of sites. Site-specific estimates which are more unreliable (i.e., based on smaller sample size), are shrunk towards the estimated population mean, resulting in more precise estimation. The empirical Bayes estimates will generally have smaller mean squared error than their OLS counterparts (Raudenbush & Bryk, 1986).

- what the average effect would have been if all eligible individuals in the demonstration site(s) had participated;
- what the effect would be if all eligible individuals in the nation participated.

The third of these, which is the hardest to measure, is the one of greatest policy interest.

If the treatment can be assumed to have the same effect on all people in all sites, then there is no need to distinguish among these interpretations; the observed treatment-control difference of *any* randomly assigned group of participants, selected in any manner, will generate the desired information. We usually suppose, however, that effects will vary along two dimensions: with the characteristics of people who participate, and with the characteristics of the sites. Those who participate in the demonstration may differ from the population of eligibles in the demonstration sites for several reasons:

- they are self-selected -- i.e., they volunteered for a chance to get into the program;
- they are selected by the program -- i.e., they are either actively recruited or screened for desirable characteristics; or
- they are selected by the evaluator -- i.e., some subgroups are oversampled because it is desirable to estimate effects separately for blacks, individuals with low income, low birthweight babies, etc.

If only the third type of selection is present, it can readily be dealt with by sample weights, because the analysis sample is still a probability sample of the eligible population. The first two types of selection, however, are based on unmeasured and probably unmeasurable characteristics. If the effects of the treatment vary with clients' motivation, for example, it may be impossible to generalize from a classic experiment performed on volunteers to the eligible

population as a whole. Thus, depending on the method of sample selection, even within a site the effect of a program may not be knowable for the entire eligible population.

The problems are multiplied when it is desired to generalize to the nation as a whole. In some studies, this is a straightforward exercise because the sites are actually a probability sample of all sites in the nation. The evaluation of the Food Stamp Employment and Training (E&T) Program (Puma et al., 1990), for example, was based on a sample of recipients in 60 local Food Stamp Agencies, which were selected based on probabilities proportional to the estimated size of their E&T caseload. Much more usual, however, is a situation in which the selected locales are a sample of convenience: chosen for their exemplary programs, or to ensure geographic diversity, or according to some other nonrandom criterion. Lacking a probability sample of sites, let us suppose that we have satisfactorily determined both what the effect of the treatment was on participants in each site, and what it would have been if all eligibles in each site had participated (or have accepted the former as an estimate of the latter). Let us further suppose that the estimates of impacts vary substantially among sites--not because of measurable differences in participant populations between sites (which can be dealt with by multivariate methods), but because of different program implementation or different environments (e.g., the local labor markets). How should we combine the site-level estimates in order to get our best guess at what the impact of nationwide implementation would be?

If we have many sites, and have good measures of site characteristics, we may be willing to model the effect size as a function of these characteristics, as described above in the HLM approach. We can then project national effects based on the national distribution of site characteristics.

This is rarely possible for social science experiments, however, for two reasons: typically too few sites participate for developing a model of site-level effects; and we often believe that the important differences in treatment effects across sites are due to site characteristics that cannot be known for the rest of the country--e.g., how the program was/would be implemented. In these situations we must combine our site-level estimates without modelling. Some criteria we might use to combine them, and their rationales, are as follows:

- Count each site-level effect equally. We have done an experiment J times, where J is the number of sites, and each of our J results is equally valid and likely to occur again.

- Count those site-level effects more heavily that are measured with more precision. We have more information about what happened in some of the experimental settings than others, and we can minimize the variance of the mean by weighting inversely proportional to the site-level variances. In the absence of multivariate analysis within each site, if the variance is constant, then this amounts to weighting by the sample size in each site. It may be, however, that the outcome has a tighter distribution in some sites than in others, or the multivariate model provides more explanatory power in some sites than in others. In this case, precision will be measured by the site-level mean squared residual.

- Count those site-level effects more heavily that correspond to a larger eligible population. Sample designs often call for drawing equal-sized samples from sites with widely varying eligible populations. We can generalize at least up to the eligible populations of the sampled sites by counting more heavily the results from the more "important" sites.

It must be acknowledged, however, that if the selected sites are not a probability sample of all possible sites, each of these is only a "best guess" at the overall effect.

An Empirical Application

We now illustrate the use of these various methods for estimating overall effects with data from the Comprehensive Child Development Program (CCDP), a five-year national

demonstration program implemented in 21 sites across the country (St. Pierre et al., 1993). The CCDP national impact evaluation is studying a variety of outcomes among 4410 families who have been randomly assigned to program and control groups within each of the sites. Hence, treatment-control comparisons within each site yield unbiased estimates of impacts for the sample members.

CCDP is designed to lead to impacts on both children and parents. The broad range of outcomes measured in this study include:

- children's cognitive or language development, social-emotional development, and physical health and growth; and
- parents' and families' economic self-sufficiency, life management skills, childrearing attitudes and skills, and psychological and physical health.

For the purposes of this illustration, a measure of children's cognitive development at 24 months of age, called the Bayley Scales of Infant Development, is used. Because the age of children taking this test ranged from 22 to 30 months, the raw scores were converted to a set of normalized standard scores with a mean of 100 and a standard deviation of 16. The sample at this stage of the analysis was a cohort of 2,587 children. This figure represented 59% of the original evaluation sample. Attrition occurred for two reasons: A large number of children were either too old or too young to include in the analysis at the time of testing; and a number of children could not be located for the testing. Comparison of the analysis cohort on a number of background characteristics collected on all the families at baseline, however, revealed no systematic differences between program and control group children across sites. Table 1 shows the standard Bayley scores for the analysis sample by program group across the 21 sites.

Table 1. Descriptive Statistics for Bayley Scores by Program Group

	PROGRAM			CONTROL		
SITE	MEAN	S.D.	N	MEAN	S.D.	N
1	93.92	17.22	75	95.16	14.56	80
2	100.00	19.73	37	95.70	16.76	43
3	88.70	16.68	30	88.50	13.91	46
4	93.52	16.10	63	95.55	16.98	82
5	98.60	17.06	52	100.98	15.75	52
6	95.39	15.68	66	93.70	14.94	66
7	94.50	16.50	26	84.26	10.79	19
8	86.93	18.05	55	84.23	16.29	66
9	93.84	18.79	44	99.66	17.16	47
10	116.18	14.55	91	115.37	18.52	88
11	104.99	18.59	71	99.53	16.25	58
12	99.09	15.96	70	91.06	20.09	68
13	86.90	18.60	59	89.11	14.09	63
14	100.53	17.91	47	95.17	19.30	69
15	111.33	21.62	57	108.01	19.12	70
16	89.27	13.95	82	89.09	13.86	79
17	99.81	13.68	63	98.31	14.72	88
18	95.72	13.57	32	101.38	19.01	39
19	98.12	18.18	64	98.76	22.29	72
20	97.69	16.43	68	95.72	17.01	68
21	105.66	20.36	83	101.74	17.59	89

Two basic approaches for computing an overall estimate of the impact of CCDP were explored. The first approach was to pool data from the 21 sites into a single analysis, comparing the overall mean outcome values on the Bayley for all program and control group children irrespective of site (Model 1). This difference was tested in an ordinary least squares

regression (OLS) model which included eight baseline covariates to increase the precision of the overall impact estimate.[4] The model underlying this approach is denoted as follows:

$$Y_i = \beta_0 + \beta_1 P_i + \beta_{2,1} X_{1,i} + \ldots + \beta_{2,8} X_{8,i} + \epsilon_i \qquad (3)$$

where

Y_i is a Bayley score for child i;

P_i is the program indicator for child i (1=Program participant, 0=Control);

X_{ki} are baseline characteristics of child i (i.e., measured prior to participation in CCDP) for $k = 1 \ldots 8$ covariates;

β's are parameters to be estimated; and

ϵ_i represents a random error term for child i.

The second approach (Model 2) estimated the overall program impact by averaging separately derived site-level impacts, using a two-stage estimation strategy. In the first stage, a single OLS regression was performed for each outcome variable, using a model which included an intercept, 8 baseline covariates, 20 site-level variables and 21 site-by-treatment interaction variables. The model was thus of the form:

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_{1,1} P_{1,i} + \ldots + \beta_{1,21} P_{21,i} \\ & + \beta_{2,1} S_{1,i} + \ldots + \beta_{2,20} S_{20,i} \\ & + \beta_{3,1} X_{1,i} + \ldots + \beta_{3,8} X_{8,i} + \epsilon_i \end{aligned} \qquad (4)$$

where

---

[4]These variables include two dummy indicators for black and Hispanic ethnicity, English as a primary language, years of mother's education, family per-person income, number of pregnancy problems during the birth of the child, number of nights spent in special care after birth, and amount of drinking during pregnancy.

P_{ji} is the program indicator for site j (1=Program participant in site j, 0=all others);

S_{ji} is the indicator for site j ($j = 1...20$);

and other terms are as defined above. The stored residuals from this analysis were then squared and averaged by site, to produce a mean squared error for each of the 21 sites.

In the second stage, a correction was made for heteroscedasticity among sites by weighting each observation by the inverse of the site-specific adjusted mean square error. The adjustment consists of multiplying the mean square error for a site by (n/(n-1)), where n is the sample size for that site. This two-stage procedure produces more accurate estimates of the standard errors than ordinary least squares.

Under this second approach, we estimate overall impacts of program outcomes by combining all our data across sites into one regression analysis. This approach differs from the simple "pooled" analysis referred to earlier, however, by explicitly modeling both site-level and site by treatment effects.[5]

In order to produce an overall estimate of impact on a given outcome variable, the 21 site-level effect estimates were averaged. The average effect estimate is accompanied by the appropriate standard error based on the pooled variance across the 21 sites. Two variants of this overall estimate were computed in which the individual site-level estimates were combined (a)

---

[5]Site-specific impacts could also be derived by estimating 21 separate models, one for each site. The single-model approach was chosen to increase precision in estimating the site-level impact parameters, on the assumption that the effects of the family background characteristics employed as baseline covariates in the model do not vary across sites.

with equal weights (Model 2a); (b) with weights inversely proportional to the variances of the estimated site-level impacts (Model 2b).[6]

For comparison purposes, a separate analysis was run using hierarchical linear modeling (Model 3). In this approach, within-site slopes (outcome score on program status) are viewed as having random variation across the sample of 21 sites (Model 3a). As the main focus is on estimating the pooled within-site slope, no attempt is made to explain any variation in slopes across sites. A within-site model is first posited, which is based on individual child-level data:

$$Y_{ij} = \beta_j X_{ij} + \theta_j P_{ij} + \epsilon_{ij} \qquad (5)$$

where

Y_{ij} is a Bayley score for child i in site j;

X_{ij} represents a vector of baseline characteristics for child i in site j;

P_{ij} is a treatment indicator variable (1 = CCDP program group; 0 = control group);

β_j is a vector of coefficients capturing the relationships of the X_{ij} with the Bayley measure within site j;[7]

θ_j is the treatment effect that represents the impact of CCDP in site j; and

ϵ_{ij} is a random error term.

---

[6]Individual site-level estimates could also be combined with weights proportional to the eligible populations in the sites. Because the CCDP samples used for this evaluation were not randomly drawn from the potential eligible populations in each site, however, this weighting approach was not a viable option.

[7]These coefficients were modeled as "fixed" estimates. That is, as in the previous models, their effects were not presumed to vary randomly across sites.

A site-level model is next formulated, in which the within-site program effect parameters θ_j are modeled as random outcome variables at the second stage. The linear model assumed to underlie θ_j is denoted:

$$\theta_j = \gamma + \upsilon_j \tag{6}$$

where

θ_j is the treatment effect slope that represents the impact of CCDP in site *j*;

γ is the average regression slope across the *21* sites; and

υ_j is the unique increment to the slope associated with site *j*.[8]

These independent errors υ_j are distributed $\sim N(0,\tau)$ with τ signifying between-group or parameter variance. Evidence of slope heterogeneity across sites, i.e., $Var(\theta_j) > 0$, was then tested against H_0: $\tau = 0$ in a Chi-square test with 20 degrees of freedom. If, according to this test, no evidence of slope heterogeneity exists (slopes do not vary randomly across sites) then a new model can be refitted where the residual slope variance is constrained to zero. In other words, θ_j is still included in the model, but is constrained to have a common fixed effect across all 21 sites (Bryk & Raudenbush, 1992). In a large-sample test of this hypothesis, the Chi-square test statistic was equal to 23.91 with 20 degrees of freedom, $p=.246$, indicating that the null hypothesis could not be rejected and that sites did not show significant variability in mean slopes relating program status to Bayley outcome scores. A new model was subsequently refitted (Model 3b), whereby the program slope was fixed across sites.

---

[8]The error term in the site-level regression is a combination of the unexplained variation in the impacts across sites and the error of estimates of the θ_j 's.

Results/Conclusions

The summary results in terms of the overall Bayley impact and its accompanying standard error for the five analytic models are displayed in Table 2.

Table 2. Results from Five Analytic Models of Estimating Overall Impact on Child's Bayley Score

MODEL	IMPACT	S.E.	T	P
1. Pooled	1.816	.7134	2.55	.006[a]
2a. Unweighted averaged	1.703	.6841	2.48	.007[a]
2b. Weighted averaged	1.544	.6435	2.39	.009[a]
3a. HLM Random	1.635	.7355	2.22	.021[b]
3b. HLM Fixed	1.673	.6611	2.53	.013[b]

[a]p-value based on N-P-2 degrees of freedom (N=total sample size, P=number of predictors in the model).
[b]p-value based on J-1 degrees of freedom (J=number of sites).

What is most striking about these results is their similarity. For this particular outcome measure, choice of approach would not alter our substantive conclusion that participation in CCDP raised children's Bayley score by about 1.5 points, a small effect in terms of practical significance. Some comments, however, in terms of bias and precision are relevant. The pooled approach (Model 1) is slightly inefficient when compared to the weighted averaged approach (Model 2b) due to differences in site means on the Bayley. Moreover, if site

differences are significant, then not accounting for site in the model could affect other terms in the model through omitted-variables bias. If Hispanics or teen-age mothers are concentrated in some sites (as was indeed the case), the estimated coefficients for these characteristics will include the site effects as well as the effects of the characteristics *per se*. This is not a major consideration in the present case, however, because our primary interest is not in the influence of ethnicity or mother's age on the child's Bayley score, but rather on the effect of the treatment. Omitted variables bias is not a danger with regard to measuring the effect of the treatment, because treatment and control group families were about equally distributed across the sites. If some sites had substantially higher proportions of treatment group families, and effects varied across sites, then there would be cause for concern.

In terms of the two averaging approaches (Models 2a and 2b), the weighted approach is only slightly more efficient. This is due to the fact that although the individual site estimates are measured with varying levels of precision, the disparity in variances across sites is not all that great. Hence, there is little gain in terms of precision from differential weighting.

Both the unweighted and weighted averaged approaches yield unbiased estimates of the overall CCDP impact. If the expected value of each site-level estimate is the same, then it can easily be shown that any choice of weights, as long as they add up to 1, will produce an unbiased estimate of the overall mean effect in the population. Thus, the only difference between the unweighted and weighted approaches is in terms of efficiency.

If, on the other hand, there is significant between-site variation in effects, then we cannot ignore that source of variation in computing our standard error of the estimate of the overall treatment effect. In the extreme case, each site's individual variation is insignificant, and

the between-site variation dominates the variance term used for weighting. Because all sites have a constant between-site variation term, this amounts to an equal weighting approach. In the more general case, (i.e., in which within-site variation is significant), the greater the between-site variation, the closer we get to an equal weighting of each site's estimate.

How can we determine if the estimates of CCDP in each site are all producing the same effect? As mentioned previously, the HLM analyses showed no significant between-site variation in the effect of CCDP on children's Bayley scores. Thus fixing the treatment parameter (Model 3b) is more efficient than letting it vary randomly across sites (Model 3a). It must be noted, however, that with only 21 sites, this evaluation design has little power to detect a significant site-by-treatment interaction effect. By relaxing the significance level for the slope heterogeneity test to $p=.10$, the risk of committing a Type II error is somewhat reduced and hence the search for interaction gains strength.

There remains unanswered the question of potential variation among sites that we are currently unable to observe because the 21 CCDP sites are not a random sample of potential CCDP sites. To the extent that the 21 sites in the evaluation were selected as "best" prospects of implementing the program, we may be seriously underestimating the variation in impact among the population of potential sites, resulting in an increase in bias in our overall impact estimates. On the other hand, to the extent that sites were selected to maximize diversity in the population, then the collection of 21 sites could have more variability than a simple random sample of sites, resulting in an overestimate of intersite variation.

The above results, while illustrating the variety of approaches available for estimating the overall impact in a multisite evaluation design, are limited in scope. It could be

hypothesized that a program of the nature of CCDP can not be expected to yield strong effects for children at 24 months of age. Other outcomes measuring, for example, parent childrearing attitudes, might be more sensitive to intersite variations in the impact of CCDP. It will be necessary, thus, to look at a wider range of outcomes before concluding that the analytic approach chosen had no impact on the obtained results. This will further aid our understanding of whether the CCDP program should be viewed as a single intervention model or if there are important differences among sites in terms of how the program is implemented and impacts upon families.

We conclude, therefore, that while we can readily make recommendations for maximizing the precision of an estimate, minimizing bias is not only a matter of specifying the correct model, but also of positing to which population the results should be generalized. Although this is an issue often overlooked in evaluation research, the question of how estimates of program impact are combined across multiple sites has serious policy implications for the interpretation of results. By informing the analysis strategy with the employed sampling design, one can on both theoretical and statistical grounds better justify conclusions drawn regarding the efficacy of a particular program intervention.

References

Bloom, H.S., Orr, L., Cave, G., Bell, S.H., & Doolittle, F. (1992). *The National JTPA study: Title II-A impacts on earnings and employment at 18 months: Executive summary.* Bethesda, MD: Abt Associates, Inc.

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Cochran, W.G. & Cox, G.M. (1957). *Experimental design.* New York: John Wiley & Sons.

Fleiss, J.L. (1986). Analysis of data from multiclinic trials. *Controlled Clinical Trials, 7(4),* 267-275.

Haney, W. (1980). Units and levels of analysis in large-scale evaluations. *New Directions for Methodology of Social and Behavioral Science, 6,* 1-15.

Hedges, L.V. (1984). Advances in statistical methods for meta-analysis. In W.H. Yeaton & P.M. Wortman (Eds.), *Issues for data synthesis. New Directions for Program Evaluation, 24,* (pp. 25-43). San Francisco: Jossey-Bass.

Infant Health and Development Program (1990). Enhancing the outcomes of low-birth-weight, premature infants: A multisite, randomized trial. *Journal of the American Medical Association, 263(22),* 3035-3042.

Puma, M.J., Burstein, N.R., Merrell, K., & Silverstein, G. (1990). *Evaluation of the Food Stamp Employment and Training Program: Final report.* Bethesda, MD: Abt Associates, Inc.

Raffe, D. (1991). Assessing the impact of a decentralised initiative: The British technical and vocational education initiative. In S.W. Raudenbush & J.D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective,* (pp. 149-167). San Diego: Academic Press.

Raudenbush, S.W. & Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59,* 1-17.

Raudenbush, S.W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13(2),* 85-116.

Sinacore, J.M. & Turpin, R.S. (1991). Multiple sites in evaluation research: A survey of organizational and methodological issues. In R.S. Turpin & J.M. Sinacore (Eds.), *Multisite evaluations. New Directions for Program Evaluation, 50,* (pp. 5-19). San Francisco: Jossey-Bass.

St. Pierre, R., Goodson, B., Layzer, J., & Bernstein, L. (1993). *National impact evaluation of the Comprehensive Child Development Program: Interim report.* Cambridge, MA: Abt Associates, Inc.